Solutions

1. To account for the uncertainty in our observations , the assumption that the i.e the noise/measurement errors that contributes to the variance in the likelihood estimate are IID(independent and identically distributed) and normally distributed is sensible because as the sample size increases, as a consequence of the Central Limit Theorem, the properly normalized sum of independent random variables(the residual errors/errors in our case) tends toward a normal distribution even if the original variables themselves are not normally distributed.
   A spherical Covariance matrix implies that the off diagonal elements are 0 meaning that there is no covariance between the components/dimensions and the variance (diagonal elements) is constant and equal.

2. If the data points are not independent the likelihood can be simplified by repeatedly applying the product rule of probability (i.e Chain Rule)

$$
\begin{aligned}
&p\left(T|f,X\right)=p\left(t_1,t_2,\ldots,t_N|f,X\right)\\
&=p\left(t_1|t_2,\ldots,t_N,f,X\right)*p\left(t_2|t_3,\ldots,t_N,f,X\right)\\
&=p\left(t_1|t_2,\ldots,t_N,f,X\right)*p\left(t_2|t_3,\ldots,t_N,f,X\right)*p\left(t_3|t_4,\ldots,t_N,f,X\right)\ldots\ldots*p\left(t_N|f,X\right)\\
&=p\left(t_N|f,X\right)\prod_{i=1}^{N-1}p\left(t_i|t_{i+1},\ldots,t_{N-1},t_N,f,X\right)
\end{aligned}
$$

3. If the data points are independent and identically distributed and $\epsilon\sim\mathcal{N}(0,\sigma^2I)$ we can express the likelihood as:

$$
p\left(T|W,X\right)=\prod_{i=1}^{N}p\left(t_i|W,x_i\right)\quad\text{, where}\quad p\left(t_i|W,x_i\right)=\mathcal{N}(Wx_i,\sigma^2I)
$$

4. The assumption about the prior distribution can result in either L_1 norm or L_2 norm in the regularization term. Consider the two cases below:
   CASE 1: Normally Distributed Priors.
   Lets assume spherical Gaussian prior with 0 mean.By applying Bayes rule, we have that

$$
\begin{aligned}
&p(\mathbf{w}|\mathcal{D})=\frac{p(\mathcal{D}|\mathbf{w})\;p(\mathbf{w})}{p(\mathcal{D})}\\
&\propto p(\mathcal{D}|\mathbf{w})\;p(\mathbf{w})\\
&\propto\Big[\prod_{n}^{N}\mathcal{N}(t^{(n)};f_{\mathbf{w}}(\mathbf{x}^{(n)}),\sigma_t^2)\Big]\,\mathcal{N}(\mathbf{w};\mathbf{0},\sigma_{\mathbf{w}}^2\mathbb{I})\\
&\propto\prod_{n}^{N}\mathcal{N}(t^{(n)};f_{\mathbf{w}}(\mathbf{x}^{(n)}),\sigma_t^2)\prod_{i=1}^{K}\mathcal{N}(w_i;\,0,\,\sigma_{\mathbf{w}}^2)
\end{aligned}
$$

   Taking negative log of probability

$$
\begin{aligned}
&-\log\left[p(\mathbf{w}|\mathcal{D})\right]=-\sum_{n=1}^{N}\log\left[\mathcal{N}(t^{(n)};f_{\mathbf{w}}(\mathbf{x}^{(n)}),\sigma_t^2)\right]-\sum_{i=1}^{K}\log\left[\mathcal{N}(w_i;\,0,\,\sigma_{\mathbf{w}}^2)\right]+const.\\
&=\frac{1}{2\sigma_t^2}\sum_{n=1}^{N}\left(t^{(n)}-f_{\mathbf{w}}(\mathbf{x}^{(n)})\right)^2+\frac{1}{2\sigma_{\mathbf{w}}^2}\sum_{i=1}^{K}w_i^2+const.
\end{aligned}
$$

In this case ,we get L$_2$ norm in regularization part (a.k.a Ridge Regression) ,the loss function is:

$$L = \underbrace{\left[\sum_{n=1}^{N}(t^{(n)} - f_{\mathbf{w}}(\mathbf{x}^{(n)}))^2\right]}_{RSS} + \underbrace{\lambda\sum_{i=1}^{K} w_i^2}_{ShrinkagePenalty}$$

CASE 2: Laplacean Priors

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})}$$

$$\propto p(\mathcal{D}|\mathbf{w})\, p(\mathbf{w})$$

$$\propto \prod_{n}^{N}\mathcal{N}(t^{(n)}; f_{\mathbf{w}}(\mathbf{x}^{(n)}), \sigma_t^2)\prod_{i=1}^{K} Laplace(w_i;\, \mu,\, b) \qquad where, \quad Laplace(\mu, b) = \frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$$

Following the same procedure as in CASE 1 , we get L$_1$ norm in regularization part (a.k.a Lasso Regression) ,the loss function is:

$$L = \underbrace{\left[\sum_{n=1}^{N}(t^{(n)} - f_{\mathbf{w}}(\mathbf{x}^{(n)}))^2\right]}_{RSS} + \underbrace{\lambda\sum_{i=1}^{K} |w_i|}_{ShrinkagePenalty}$$

The shrinkage penalty is a regularization term.As illustrated in the figure below from Bishop's Pattern Recognition and Machine Learning- the intersection of the contour plots for the un-regularised and the regularized loss function is more likely to result in sparse solutions in case of L$_1$ norm. In other words, the L$_1$ norm can result in a sparse model(where some parameters become 0) resulting in variable selection. This is because the Laplace distribution has a sharper peak.

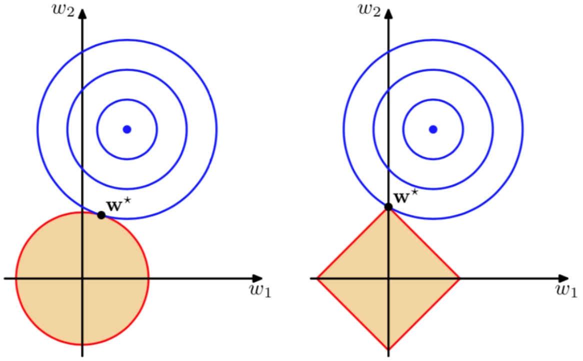


Figure 1: Contour of Unregularized Error Function(Blue) along with constraint region. L2 Norm on left and L1 norm on the right

5. We know that the target variables in t are mentioned to be conditionally independent with $p\,(t_i|W_i, x) = \mathcal{N}(XW, \sigma^2 I)$ and $p(\mathbf{w_i}) = \mathcal{N}(\mathbf{w_i}; w_0, \tau_{\mathbf{w}}^2\mathbb{I})$.By using these assumptions and applying Bayes rule:

$$p(W_i|\mathbf{X}, T_i) = \frac{p(T_i|\mathbf{X}, W_i)\, p(\mathbf{w})}{z}$$

$$\propto p(\mathbf{T_i}|\mathbf{X}, \mathbf{W_i})\, p(\mathbf{w})$$

Since the likelihood and the prior is Gaussian,due to the conjugate property, the resulting posterior will also be Gaussian

$$\propto \Big[\prod_{n}^{N} \mathcal{N}(t^{(n)}; f_{\mathbf{w}}(\mathbf{x}^{(n)}), \sigma_t^2 \mathbb{I})\Big]\, \mathcal{N}(\mathbf{w}; w_0, \tau_{\mathbf{w}}^2 \mathbb{I})$$

LHS
Let the mean for the posterior be μ_w and the variance be Σ_W. The exponents for the left hand side will be

$$\frac{-1}{2} W^T \Sigma_w^{-1} W + W^T \Sigma_w^{-1} \mu_w + \frac{-1}{2} \mu_w \Sigma_w^{-1} \mu_w$$

RHS
The exponents for the right hand side are mentioned below

$$-\frac{1}{2\sigma^2}(XW)^T(XW) + \frac{1}{\sigma^2} t^T(XW) + \frac{-1}{2\sigma^2} t^T t + \frac{-1}{2\tau^2} W^T W + \frac{1}{\tau^2} W^T W_0 + \frac{-1}{2\tau^2} W_0^T W_0$$

By completing the square on LHS and RHS we get the below results

$$\Sigma_W^{-1} = \frac{1}{\tau^2} + \frac{X^T X}{\sigma^2}$$

$$\Sigma_W = \Big(\frac{1}{\tau^2} + \frac{X^T X}{\sigma^2}\Big)^{-1}$$

$$\mu_W = \Big(\frac{1}{\tau^2} + \frac{X^T X}{\sigma^2}\Big)^{-1} \Big(\frac{X^T t}{\sigma^2} + \frac{W_0}{\tau^2}\Big)$$

From Bayes Theorem we have

$$Posterior = \frac{Likelihood * Prior}{Evidence}$$

$$\implies z = likelihood$$

which is just a normalization term to ensure that the resulting probability distribution function is valid i.e. sums to 1.
In likelihood estimate , we can estimate model parameters based only on the minimization of sum of squared error, $W_{MLE} = (X^T X)^{-1} X^T t$.Once we have W_{MLE}, we can make predictions for t^*. Where as in Bayesian Linear Regression,we calculate the posterior distribution. We express our beliefs about a prior. This prior term introduces an additional term to the loss function which is a regularization term. The equations are written in question 4 for reference.

6. Gaussian process is a prior over function realisations.From the definition of a Gaussian process, the marginal distribution $p(f|X,\theta)$ is given by a Gaussian whose mean is zero and whose covariance is defined by a Gram matrix K so that $p(f|X,\theta) = N(f|0, K)$. Here K is a kernel function which is chosen to express a property that, for points x_n and x_m that are similar, the corresponding values $f(x_n)$ and $f(x_m)$ will be more strongly correlated than for dissimilar points.
Since we are not interested directly in f , we can can marginalise it out

$$p(\mathbf{T}|\mathbf{X},\theta) = \int \underbrace{p(\mathbf{T}|f)}_{likelihood} * \underbrace{p(f|\mathbf{X},\theta)}_{prior} df$$

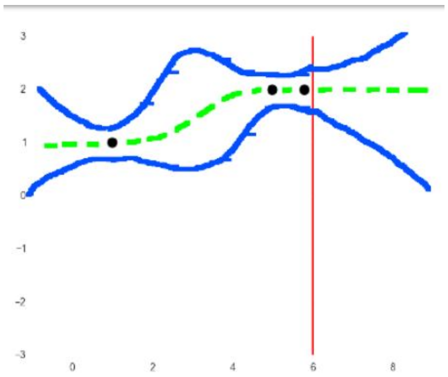


Figure 2: Effect of prior in GP

The effect of the prior over functions as we observe data can be seen from the figure 2.Predictive uncertainty is smallest in the neighbourhood of the data points.The marginalization of prior over functions and the data is filtering out the uncertainty.

7. Assuming X and θ are independent and target variable $\mathbf{t}$ is conditionally independent of X and θ given f , we have the below graphical model.

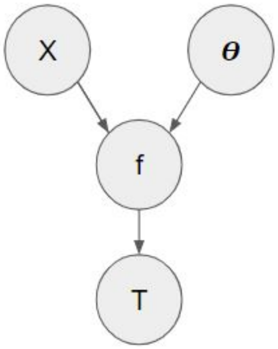


Figure 3: Graphical Model for $p(\mathbf{T}, \mathbf{X}, f, \theta)$

The joint likelihood of the full model can be written as follows:

$$p(\mathbf{T}, \mathbf{X}, f, \theta) = p(\mathbf{T}|f) * p(f|\mathbf{X}, \theta) * p(\mathbf{X}) * p(\theta)$$

8. The required marginalization is shown below:

$$p(\mathbf{T}|\mathbf{X}, \theta) = \int p(\mathbf{T}|f) * p(f|\mathbf{X}, \theta) df$$

- We have modeled the relationship between T and f and also f and X but we really are interested in the relationship between T and X.We are not interested in f and therefore the variable is marginalised. We average the likelihood of data points over all the possible functions given by the prior over function realisations.
- There are two sources of uncertainty , one which is associated with our beliefs of the functions, and the other, in how we believe the output of the function have generated the observed data. These two sources are independent and as a result are merged i.e simply added to form the covaraince of the marginal likelihood.The level of uncertainty decreases as more data points are observed.This can be seen in figure 2.

- θ is the hyper-parameter which comes from the kernel function.When averaging over function realisations of GP, it is constant in the integral so it still remains on the left side of the expression.

9. .

1 The prior distribution over W is assumed to Gaussian with the mean and covariance mentioned below.
$\mu = [0,0] \quad \Sigma = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$

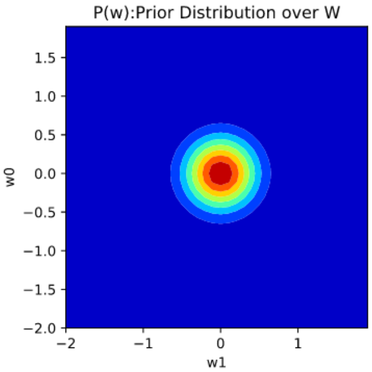


Figure 4: Prior Distribution over W,$p(\mathbf{w})$

2 Figure 5(a) shows the posterior distribution over $\mathbf{W}$ after observing 1 data point.

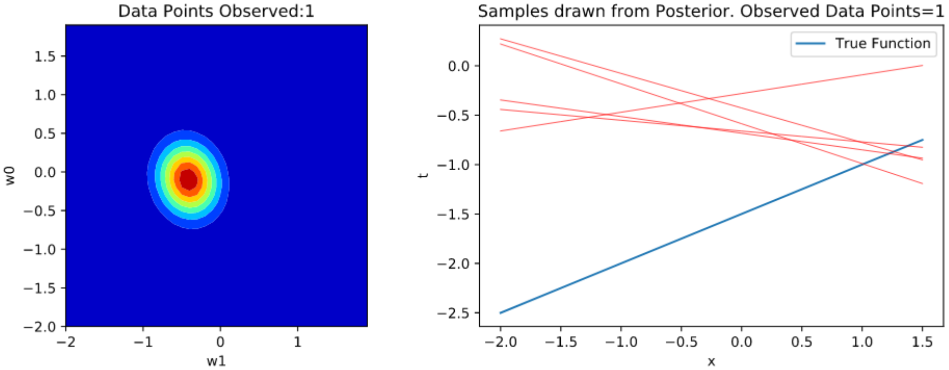


Figure 5: Posterior distribution over W (left) and samples of functions drawn (right) after observing 1 data point.

3 Figure 6(b) shows the plot of the resulting functions after drawing 5 sample data points from the posterior distribution over W having observed data point.

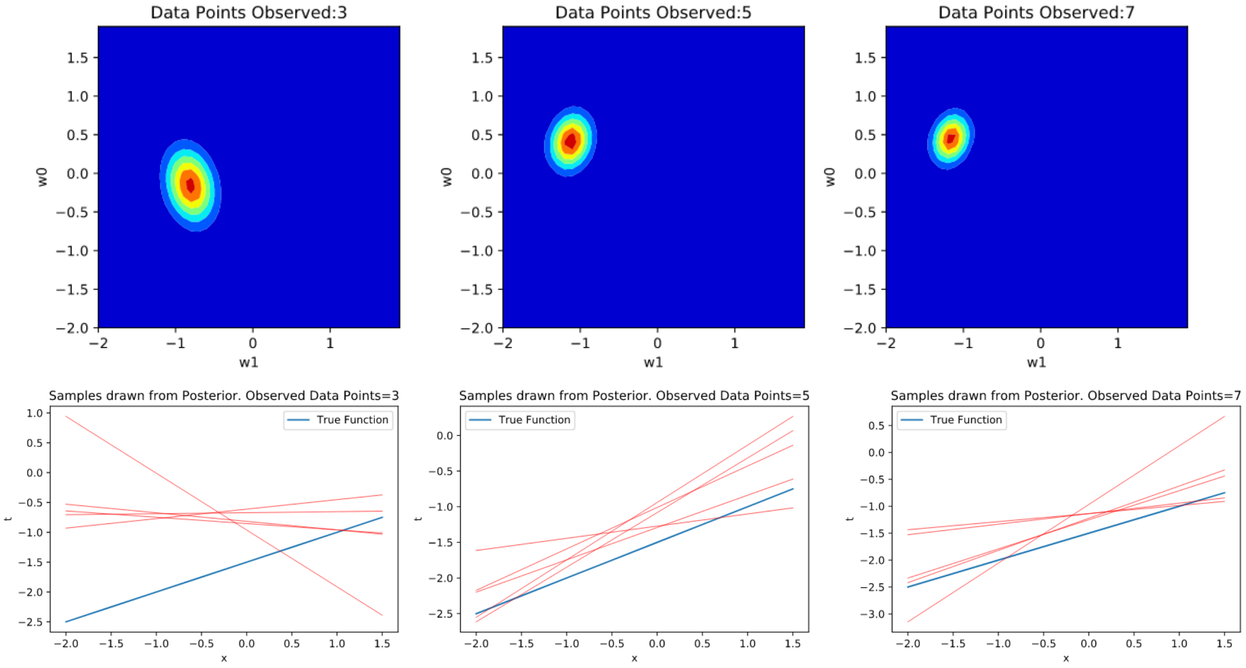


Figure 6: Posterior distribution over W (top) and samples of functions drawn (bottom)

4 Figure 6 show the posterior distribution over W(top) and plot of the resulting functions (after drawing 5 sample data points)(bottom) after observing [3,5,7] data points.

5 After observing more data points , it can clearly be seen in figure 7 that the posterior(left) is becoming narrower and and plot of the resulting functions(right) sampled from the posterior are getting close to the underlying true function. The posterior distribution over W is centered / getting close to the true values of $W_0 = 0.5$ and $w_1 = -1.5$ as number of data points is increasing. See figure 7 where observed data points is 100.

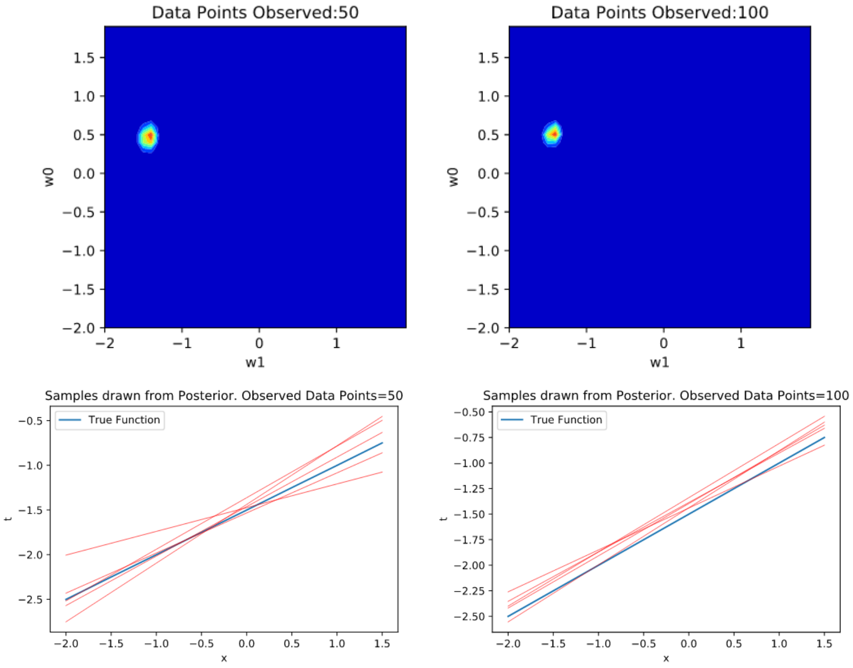


Figure 7: Posterior distribution over W (top) and samples of functions drawn (bottom)

6 Figure 8 shows the effect of noise on the posterior distribution over W. The results are as expected, with the increase in the level of noise, the uncertainty is increasing.

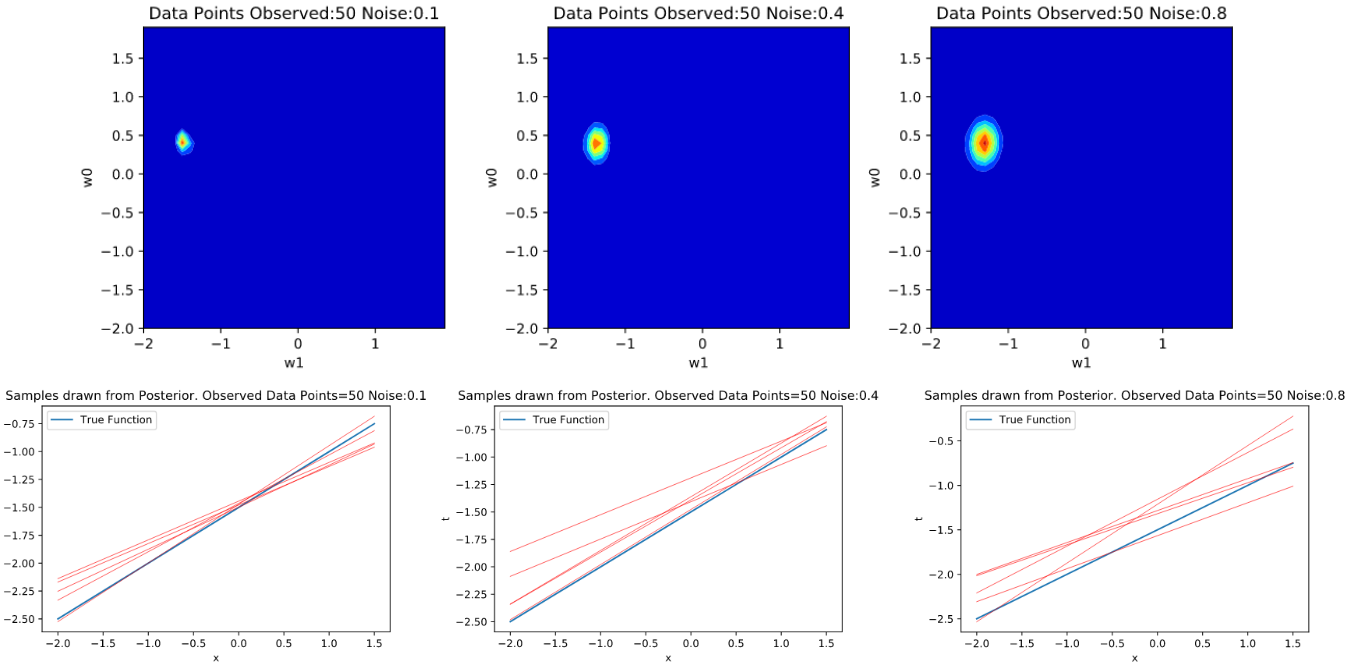


Figure 8: Effect of noise levels on the posterior

10. .

1 Figure 9 shows some samples drawn from a Gaussian prior with a squared exponential covariance function.

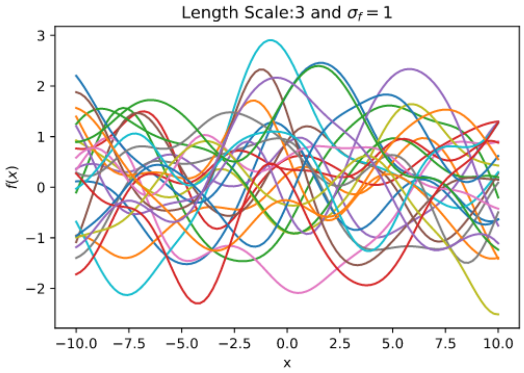


Figure 9: Samples drawn from a GP-prior with a squared exponential covariance function

2 The figures below show 10 samples drawn from a GP Prior with varying length scales. The same value of $\sigma_f = 1$ was used to produce the plots. The kernel function that determines K is typically chosen to express the property that, for points x_n and x_m that are similar, the corresponding values $y(x_n)$ and $y(x_m)$ will be more strongly correlated than for dissimilar points . In the function $k(x_i, x_j) = \sigma_f^2 e - \frac{(x_i - x_j)^T (x_i - x_j)}{l^2}$ It is clearly evident that as the value of length scale increases the function realisations become smoother. For larger L values the function instantiations are more strongly related and the function looks smoother as there is more dependency between the locations, and for small length scale values, there is little dependency between the points so the function looks more wiggly.

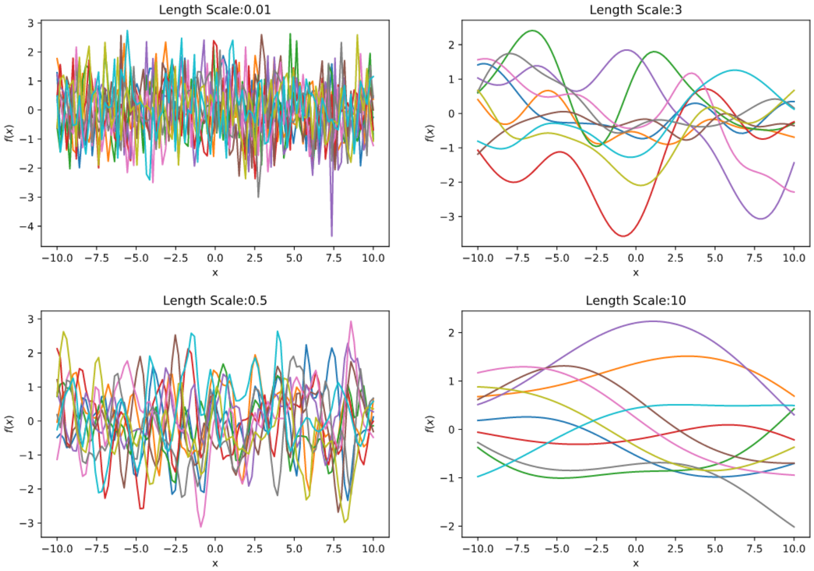


Figure 10: 10 Samples drawn from a Gaussian Prior with different Length scales

11. 1 The prior allow us to formulate our uncertainty in our beliefs. We combine this prior with our observations/data to compute the posterior distribution that facilitates our learning. In absence of data or observations we can say that the prior and the posterior distributions is the same object.

2 We want to make predictions t_* for n_2 new samples after observing n_1 data points (x_{obs}, t_{obs}). We are interested in computing the posterior $p(\mathbf{t}_* \mid \mathbf{t_{obs}}, X_{obs}, X_*)$. Due to conjugate property of Gaussian's the conditional and marginal distributions shown below will also be Gaussian.

$$\begin{bmatrix} \mathbf{t}_{obs} \\ \mathbf{t}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

where ,

$$\mu_1 = m(X_{obs}) \quad (n_1 \times 1)$$
$$\mu_2 = m(X_*) \quad (n_2 \times 1)$$
$$\Sigma_{11} = k(X_{obs}, X_{obs}) + \sigma_\epsilon^2 I \quad (n_1 \times n_1)$$
$$\Sigma_{22} = k(X_*, X_*) \quad (n_2 \times n_2)$$
$$\Sigma_{12} = k(X_{obs}, X_*) = k_{21}^\top \quad (n_1 \times n_2)$$

The conditional distribution can be given as:

$$p(\mathbf{t}_* \mid \mathbf{t}_{obs}, X_{obs}, X_*) = \mathcal{N}(\mu_{2|1}, \Sigma_{2|1})$$

where,

$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{t}_{obs} - \mu_1)$$
$$= \Sigma_{21}\Sigma_{11}^{-1}\mathbf{t}_{obs} \quad \text{(if assume mean prior } \mu = 0)$$
$$= (\Sigma_{11}^{-1}\Sigma_{12})^\top \mathbf{t}_{obs}$$

and ,

$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$
$$= \Sigma_{22} - (\Sigma_{11}^{-1}\Sigma_{12})^\top \Sigma_{12}$$

Figure 11 shows the predictive posterior distribution for the model with two length scales.

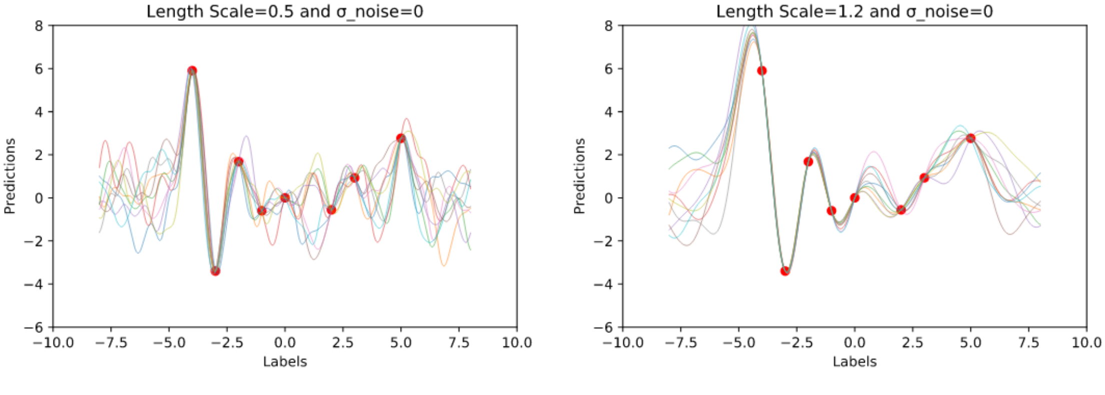


Figure 11: Samples from posterior distribution with 8000 points between [-8,8]

3 Some samples from the posterior distribution with points close to and far away from the observed data are drawn and plotted for two different length scale in Figure 11. It can be observed that the all samples in below cases pass through the observed points since $\sigma_{noise} = 0$. We also observe that the uncertainty is high in regions/areas where we don't observe data(towards the left and right most region) and it decreases in regions close to the observed data points.

4 Figure 12 shows the data,predictive mean and the predictive variance from the posterior distribution for two different length scales.

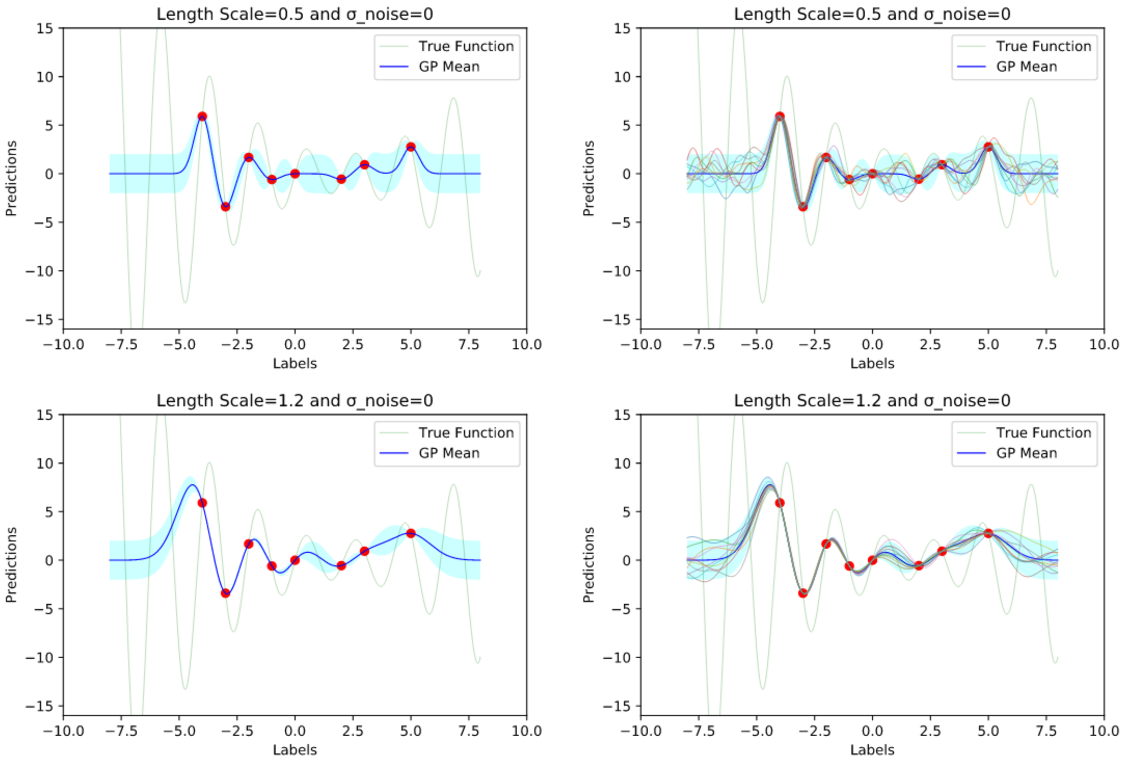


Figure 12: Plot of predictive μ and $\mu \pm 2\sigma$, data points and samples(right) from posterior distribution

5 As mentioned before the prior allow us to formulate our uncertainty in our beliefs. We combine this prior with our observations/data to compute the posterior distribution that facilitates our learning.The uncertainty decreases in regions where we observe more data and is still high in regions away from observed data points.(figure 12 show this).

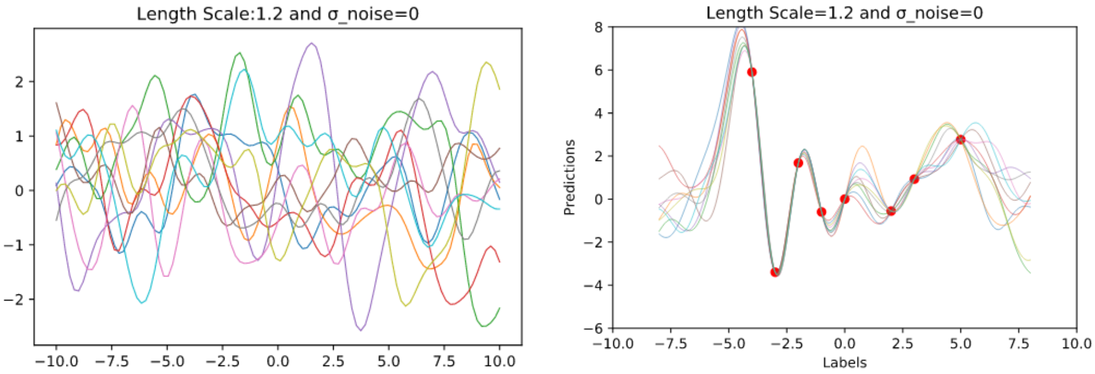


Figure 13: Some samples drawn from Prior(left) and Posterior Distribution(right)

6 So far we have assumed a noiseless distribution. This is evident from the Figures 12,13 since the samples drawn

and their predictive mean all pass through the observed data points i.e posterior variance becomes 0 at those points.In presence of white noise (iid), there is an additional term added to the kernel matrix as shown below

$$\Sigma_{11} = k(X_{obs}, X_{obs}) + \sigma_\epsilon^2 I$$

As a result , the predictive mean,and the samples from the posterior distribution no longer pass from the observed points and are corrupted due to this noise. The posterior variance at these data points is no longer 0.

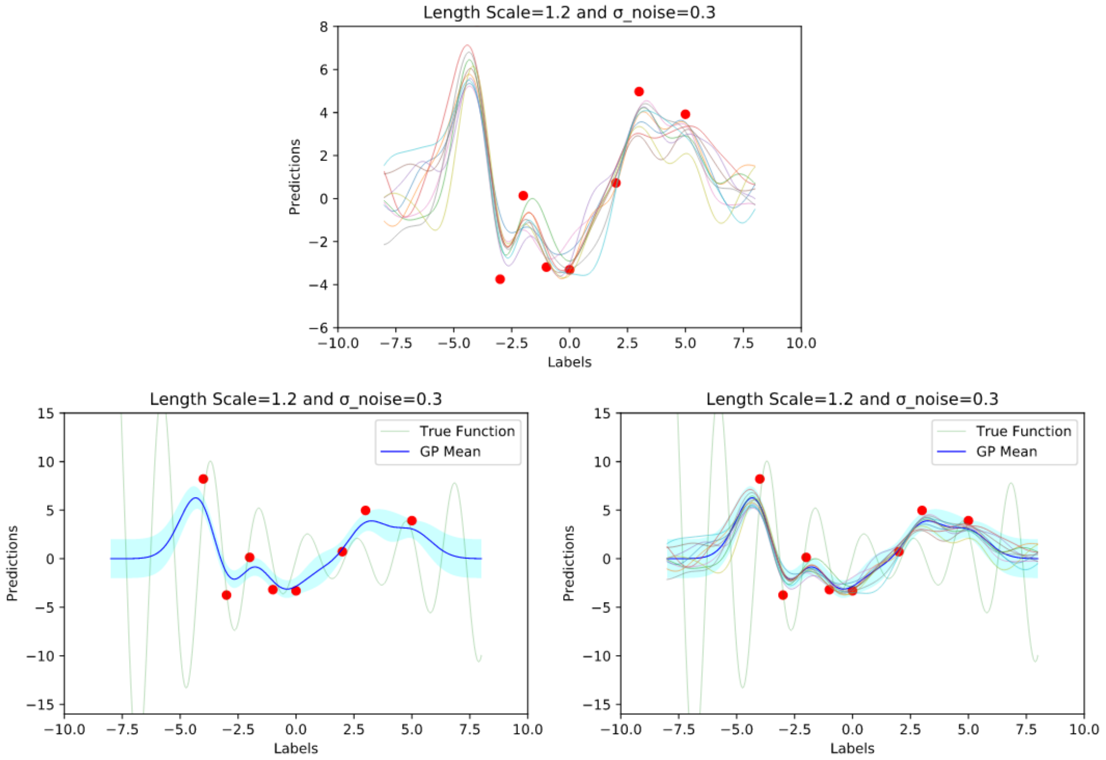


Figure 14: Plot of predictive μ and $\mu \pm 2\sigma$, data points and samples(right) from posterior distribution with diagnal noise

12. The prior expresses our beliefs/preferences on the latent variable X. With the assumption

$$p(X) \sim \mathcal{N}(0, I)$$

the following preferences are encoded:

- Underlying latent variable X has a Gaussian distribution with 0 mean.
- Different dimensions of X are independent of each other(not co related).

13.

$$y = Wx + \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

We can marginalize over the latent variables

$$p(y|W) = \int p(y|x, W)p(x)dx \quad \text{where} \quad p(y|x, W) = \mathcal{N}(y|Wx, \sigma^2 I) \quad \textit{and} \quad p(x) = \mathcal{N}(x|0, I)$$

Due to the conjugate property of gaussians we know that the resulting marginla distribution is also Gaussian i.e We can compute the mean and covariance given that it is Gaussian

$$\begin{aligned}\mathbb{E}[y|W] &= \mathbb{E}[Wx+\epsilon]\\ &= W\mathbb{E}[x]+\mathbb{E}[\epsilon]\\ &= W*0+0\\ &= 0\end{aligned}$$

$$\begin{aligned}C &= \mathbb{E}[(y_i-\mathbb{E}[y])(y-\mathbb{E}[y])^T]\\ &= \mathbb{E}[(Wx+\mu+\epsilon-\mu)(Wx+\mu+\epsilon-\mu)^T]\\ &= \mathbb{E}[(Wx+\epsilon)(Wx+\epsilon)^T]\\ &= WW^T+\sigma^2 I\end{aligned}$$

$$p(y_i|W) = \mathcal{N}(0, WW^T+\sigma^2 I)$$

14. RESULTS ARE FOR NEGATIVE LOG SPACE

MLE Estimate In this case maximising the likelihood is equivalent to minimizing the sum squared error.

$$L(W) = \frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_n - w^T x_n)^2 + const$$

MAP Estimate For MAP estimate , we calculate the posterior. The second term in the below MAP estimate is a result of the prior.(derived in question 4 for a gaussian prior) This acts as a regularization term.

$$L(W) = \frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_n - w^T x_n)^2 + \frac{\lambda}{2}(w^T w) + const$$

1 As per the the Bernstein–von Mises theorem, the posterior distribution for unknown quantities in any problem is effectively asymptotically independent of the prior distribution (assuming it obeys Cromwell's rule) as the data sample grows large. The data gets precedence over the prior and the MAP estimate converges to the MLE estimate.

2 When we are looking at the optimization problem here i.e to estimate optimal parameter W.The denominator is independent of W ,we can optimize the below equation. This will give the same result.

$$argmax_W \ \frac{p(Y|X,W)p(W)}{\int p(Y|X,W)p(W)dW} \propto argmax_W \ p(Y|X,W)p(W)$$

3 **TYPE 2 Maximum Likelihood**
In the representation task we have a model with two latent variables W and X that interact. This means that a Type-II ML estimation is a sensible approach to learn the model i.e we integrate out one parameter (X) and then maximise over another(W), as follows:

$$\hat{W} = argmax_{\mathbf{W}} \int p(Y|X,W)p(X)dX$$

15. .

1 From Q14 we have , the following

$$p(Y|W) = \prod_{i=1}^{N} \mathcal{N}(y_i|0, WW^T + \sigma^2 I)$$

then the objective function is derived as follows using log rules:

$$\begin{aligned}
\mathcal{L}(W) &= -\log\left(\prod_{i=1}^{N} \mathcal{N}(y_i|0, WW^T + \sigma^2 I)\right) \\
&= -\sum_{i=1}^{N} \log\left(\mathcal{N}(y_i|0, WW^T + \sigma^2 I)\right) \\
&= -\sum_{i=1}^{N} \log\left(\frac{1}{2\pi^{\frac{D}{2}}\sqrt{|\Sigma|}}\right) - \sum_{i=1}^{N} \log(e^{\left(-\frac{1}{2}y_i^T\Sigma^{-1}y_i\right)}) \\
&= -\sum_{i=1}^{N} \log\frac{1}{(2\pi)^{\frac{D}{2}}|WW^T + \sigma^2 I|^{\frac{1}{2}}} - \sum_{i=1}^{N} \log(e^{\left(-\frac{1}{2}y_i^T(WW^T+\sigma^2 I)^{-1}y_i\right)}) \\
&= \sum_{i=1}^{N} \log(2\pi)^{\frac{D}{2}}|WW^T + \sigma^2 I|^{\frac{1}{2}} + \sum_{i=1}^{N} \frac{1}{2} y_i^T (WW^T + \sigma^2 I)^{-1} y_i \\
&= \frac{N}{2}\left(D\log 2\pi + \log(|WW^T + \sigma^2 I|)\right) + \frac{1}{2}\sum_{i=1}^{N} y_i^T (WW^T + \sigma^2 I)^{-1} y_i
\end{aligned}$$

2 The negative log likelihood derived previously can also be written as

$$\mathcal{L}(W) = \frac{N}{2}\big(\overbrace{D\log 2\pi}^{Term1} + \overbrace{\log(|WW^T + \sigma^2 I|)}^{Term2} + \overbrace{Tr\big((WW^T + \sigma^2 I)^{-1} YY^T\big)}^{Term3}\big)$$

Derivative for Term 1 The first term is independent of W so we can ignore it.
Derivative for Term 2
From matrix cookbook $\partial(\log(\det(X))) = (X^{-1}\partial X)$,also $\sigma^2 I$ is constant w.r.t to W :

$$\frac{\partial Term2}{\partial W} = \Big((WW^T + \sigma^2 I)^{-1}\partial(WW^T)\Big)$$

Also, $\partial(XY) = (\partial X)Y + X(\partial Y)$ and that $\frac{\partial X}{\partial X_{ij}} = J^{ij}$ where J is the single-entry matrix, having 1 at (i, j) and 0 elsewhere.

$$\frac{\partial Term2}{\partial W_{ij}} = \Big((WW^T + \sigma^2 I)^{-1}(J^{ij}W^T + WJ^{ijT})\Big)$$

Derivative for Term 3
Using $\partial(Tr(X)) = Tr(\partial X)$ and previous formulas we have:

$$\frac{\partial Term3}{\partial W} = Tr\Big(\partial\Big(Y(WW^T+\sigma^2 I)^{-1}Y^T\Big)\Big) = Tr\left(YY^T\partial\Big((WW^T+\sigma^2 I)^{-1}\Big) + (WW^T+\sigma^2 I)^{-1}\Big(\partial(Y^TY)\Big)\right)$$

$$= Tr\left(YY^T\partial\Big((WW^T + \sigma^2 I)^{-1}\Big)\right)$$

also $\partial(X^{-1}) = -X^{-1}(\partial X)X^{-1}$ we can further simplify:

$$= Tr\left(YY^T\Big(-(WW^T+\sigma^2 I)^{-1}(J^{ij}W^T + WJ^{ijT})(WW^T+\sigma^2 I)^{-1}\Big)\right)$$

Combining all derivatives together we have

$$\frac{\partial\mathcal{L}}{\partial W_{ij}} = \frac{N}{2}Tr\left(YY^T\Big(-(WW^T+\sigma^2 I)^{-1}(J^{ij}W^T+WJ^{ijT})(WW^T+\sigma^2 I)^{-1}\Big)\right)+\frac{N}{2}Tr\Big((WW^T+\sigma^2 I)^{-1}(J^{ij}W^T+WJ^{ijT})\Big)$$

16. .

1 The figures below show the 2d representation of the learned variable X (in red). The values for N ,σ used for the plots are mentioned in the title.The estimates of W obtained from the scipy optimize method were used to learn the 2d representation of X.

$$X^* = M^{-1}W_{ML}T \quad where \quad M = (W_{ML}^T W_{ML} + \sigma^2 I)$$

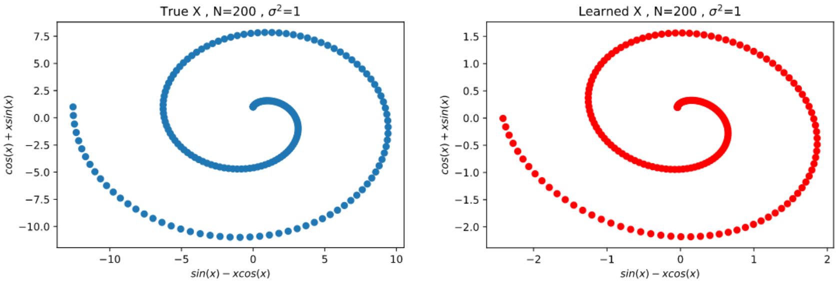


Figure 15: Representation learning for Latent Variable X

2 The learned "representation" of X resembles the shape of the true X as evidenced from the figures. However, there is some rotation in the representation of learned variable. This can be explained as follows:

$$P(Y|W) = \prod_{i=1}^{N}\mathcal{N}(y_i|0, WW^T + \sigma^2 I)$$

For all $\hat{W} = WR$ such that $RR^T = I$:

$$P(Y|\hat{W}) = \prod_{i=1}^{N}\mathcal{N}(y_i|0, WRR^TW^T + \sigma^2 I)$$
$$= \prod_{i=1}^{N}\mathcal{N}(y_i|0, WW^T + \sigma^2 I)$$

The likelihood is invariant to such rotations of W. It is also observed that the scale of learned X is different . This is because we are plotting the expected value i.e mean of the latent variable. An alternative approach is to use SVD to compute the closed form estimates of W and σ as per formulas derived in the textbook and plot the mean with standard deviations.

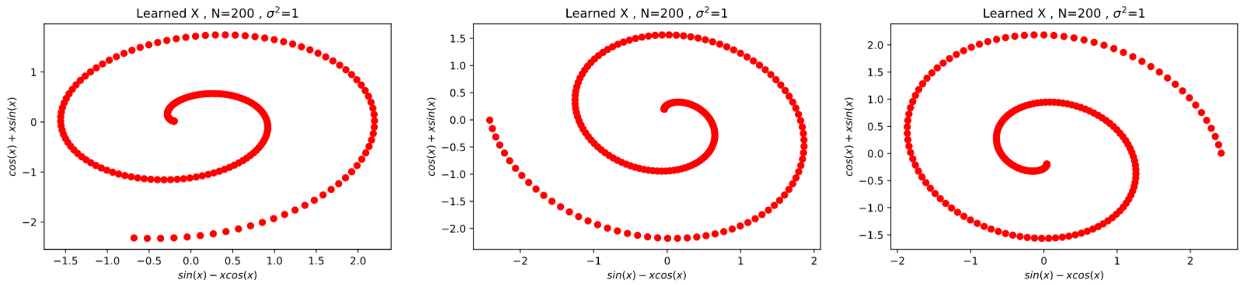


Figure 16: Invariances in Representation learning for Latent Variable X

3 The plots below show the effect of N, number of samples on representation learning for variable **X** .With decreasing values of N , the task of learning the representation of X is getting more and more difficult.

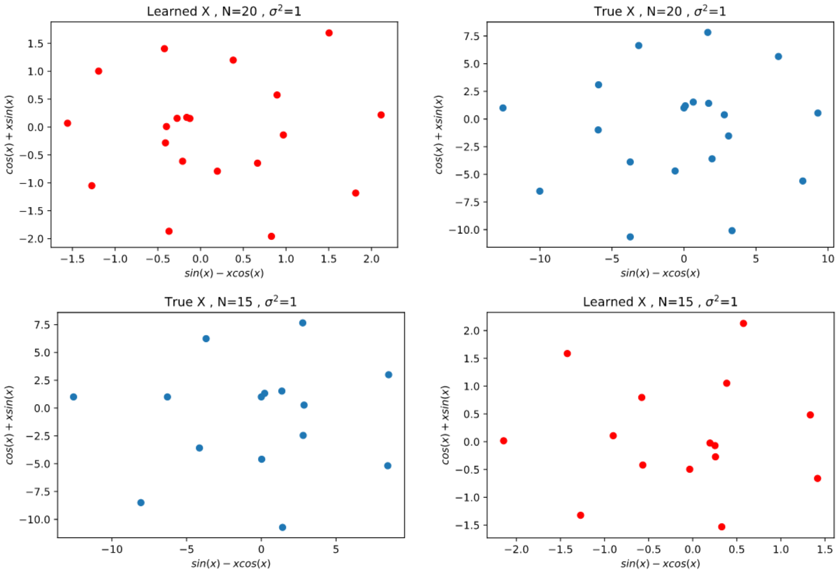


Figure 17: Effect of observed data points on representation learning

17. **MODEL 0** This is the simplest in terms of parameter counting, it has no free parameters.It treats all data sets equally. This implies that it simply takes the probability mass and places it uniformly over the whole data space.

$$p(D|M_0, \theta_0) = \frac{1}{512}$$

This model could also be considered bad because it assigns all the datasets with many having different types of behavior , the same probability mass.It has no flexibility and uses no information about D except for its cardinality.

18. **MODEL 1**

$$p(D|M_1, \theta_1) = \prod_{n=1}^{9} p(t^{(n)}|M_1, \theta_1) = \frac{1}{1 + e^{-t^n \theta_1^1 x_1^n}}$$

Unlike Model 0 , that assigns a uniform probability mass to all data sets, Model 1 is a variation of logistic regression where $\theta_1^2 = 0, \theta_1^3 = 0$.From the above equation , it is evident that this model can only form decision boundaries that are a function of dimension x_1.It simply ignores the second dimension of x and the bias.It assigns more probability mass to the data sets which can be separated by a decision boundary which is a function of x_1

19. **MODEL 3** it has the most parameters and can realize the other models by setting some of its parameters to zero. This means it can spread the bulk of its unit probability mass over a wider range of data sets than the other models.It is the full logistic regression model which is flexible and can form complex decision boundaries but since it spreads out its probability mass to many data sets, it loses to simpler models.

$$p(D|M_3, \theta_3) = \frac{1}{1 + e^{-t^n(\theta_3^1 x_1^n + \theta_3^2 x_2^n + \theta_3^3)}}$$

MODEL 2 Unlike model 1 , which can only form decision boundaries that are a function of dimension x_1. This function can realize decision boundaries which are also a function of dimension x_2. In such cases it spreads more probability mass as compared to model 3 and therefore is favored. It cannot form decision boundaries that are offset from origin due to the missing bias term.

$$p(D|M_2, \theta_3) = \frac{1}{1 + e^{-t^n(\theta_2^1 x_1^n + \theta_2^2 x_2^n)}}$$

20. By being Bayesian , we need to account for the uncertainty in not only the data but also the parameters θ. We do so by expressing our belief for the model parameters by assigning it a prior . In order to compute the evidence i.e probability of the data belonging to a model we need to remove the dependency of the parameters by marginalizing them out . This can be done by the below integration.

$$p(D|M_i) = \int_{\forall\theta} p(D|M_i, \theta)p(\theta)d\theta$$

21. The choice of prior $P(\theta|M_i) \sim \mathcal{N}(0, 10^3\mathbf{I})$ implies it is a Gaussian distribution with zero mean and a diagonal covariance. This implies that the dimensions are independent and identically distributed. The high Σ results in decision boundaries which are very sharp in data space.

22. The evidence over the whole data set for each model can be seen in figure 18. The figure on the right is magnified. Some observations are:

    - For all the models, the evidence over entire **D** sums to 1. This is because they are all probability distribution functions.
    - M0 has the same evidence value for the whole D .For majority of the data set (>100), it is the model which assigns higher evidence.
    - In some regions models M1,M2 assign significantly greater probability mass as compared to M3. This is because they are simpler.
    - For few datasets M3 has highest probability mass , these could be datasets for which the decision boundary is offset from the origin because of bias term, which is only present in M3.
    - For some regions , M2 has higher probability mass than M1, because the decision boundary is a function of both x_1 and x_2. and passes through origin.

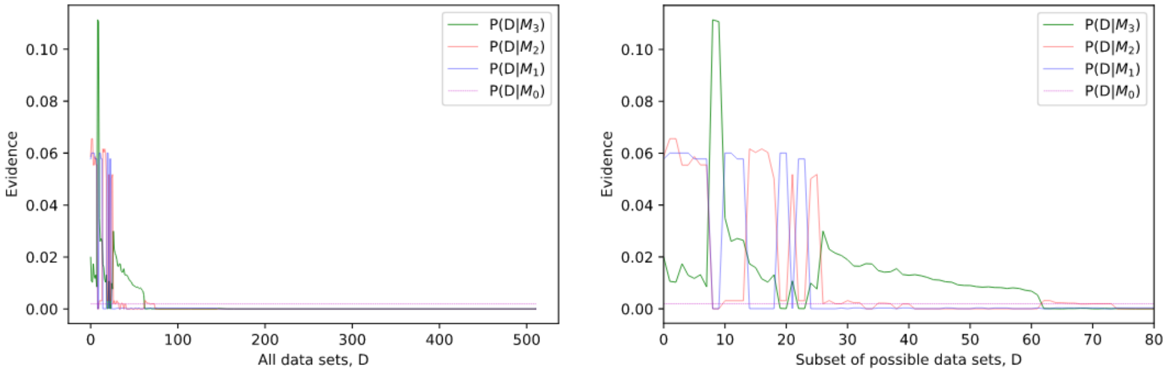
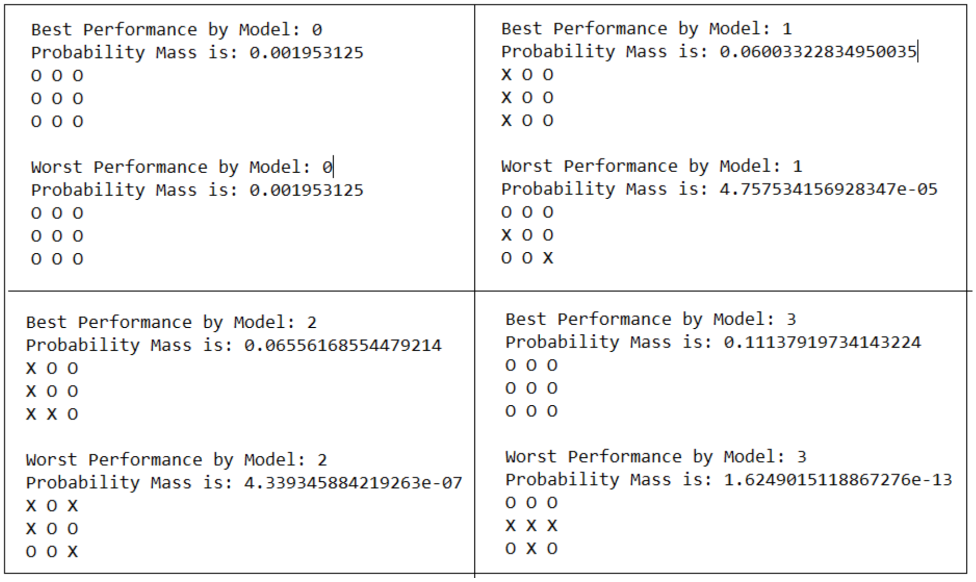


Figure 18: Plot of evidence for all possible data sets for the models, for S=1000

23. Figure 17 shows the datasets which give the highest and lowest evidence for each model along the evidence value.



```
Best Performance by Model: 0
Probability Mass is: 0.001953125
O O O
O O O
O O O

Worst Performance by Model: 0
Probability Mass is: 0.001953125
O O O
O O O
O O O
```

```
Best Performance by Model: 1
Probability Mass is: 0.06003322834950035
X O O
X O O
X O O

Worst Performance by Model: 1
Probability Mass is: 4.757534156928347e-05
O O O
X O O
O O X
```

```
Best Performance by Model: 2
Probability Mass is: 0.06556168554479214
X O O
X O O
X X O

Worst Performance by Model: 2
Probability Mass is: 4.339345884219263e-07
X O X
X O O
O O X
```

```
Best Performance by Model: 3
Probability Mass is: 0.11137919734143224
O O O
O O O
O O O

Worst Performance by Model: 3
Probability Mass is: 1.6249015118867276e-13
O O O
X X X
O X O
```

Figure 19: Highest and Lowest evidence for each model

MODEL 0 It assigns the same probability mass to each data set. This is evident from the evidence values from the figure.
MODEL 1 The dataset with highest evidence can be separated by a decision boundary which is only a function of dimension x_1 and passes through origin . The dataset with least probability mass, requires a complex decision boundary which is a function of both x_1 and x_2. Hence M1 assigns low evidence.
MODEL 2 The dataset with highest evidence can be separated by a decision boundary passing through origin and which is dependent on both x_1 and x_2 so M2 gives higher probability mass to it.The dataset with least probability mass, requires a complex decision boundary, so M2 assigns less probability mass to it.
MODEL 3 The dataset with highest evidence can only be separated by a decision boundary that is offset from origin, only M3 can account for it due to the bias term. The dataset with least probability mass cannot be well modeled by any sharp linear boundary, in this case the uniform model M0 is most likely.

24. .

- For lower variance , the models have a more uniform evidence as expected and the graph is not sharply peaked anymore. See below figure for reference. $\Sigma = I$ was used here.

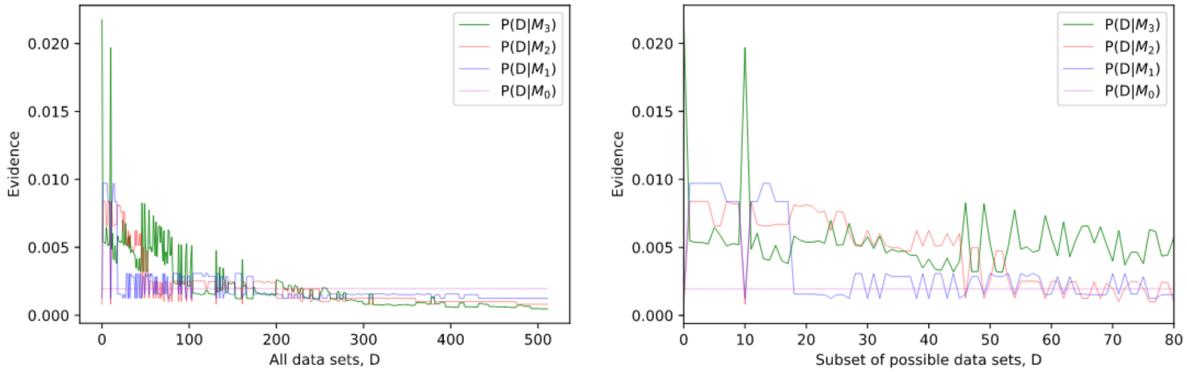


Figure 20: Plot of evidence for all possible data sets for the models, for S=1000 and low (diagonal) covariance

- For Non Zero mean, it was observed (see below figure) that the evidence is more sharply peaked for some datasets for all models. These peaks might be around the mean for each parameter. The covariance used here was the same as the problem in the assignment and mean used was 5 for each dimension.

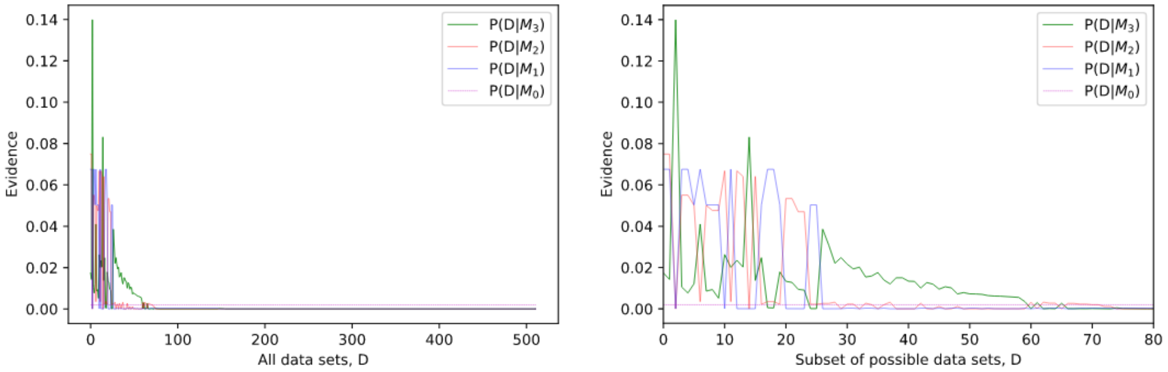


Figure 21: Plot of evidence for all possible data sets for the models, for S=1000 and non zero mean

- By using a non-diagonal covariance matrix,we introduce dependencies between different dimensions of θ. Depending on the type of co relation between these parameters , the evidence assigned by each model varies. The plot of evidence for D for each model is shown below. The positive semi definite matrix for covaraince was randomly generated.

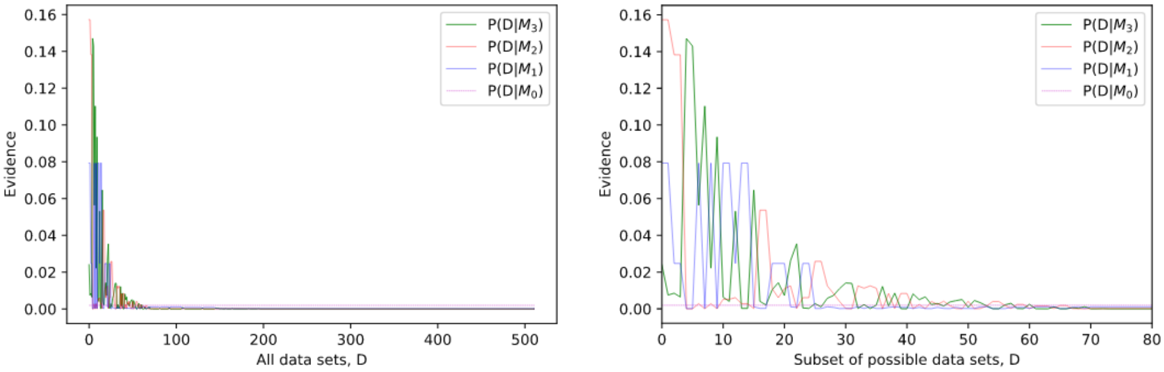


Figure 22: Plot of evidence for all possible data sets for the models, for S=1000 for non diagonal covariance matrix